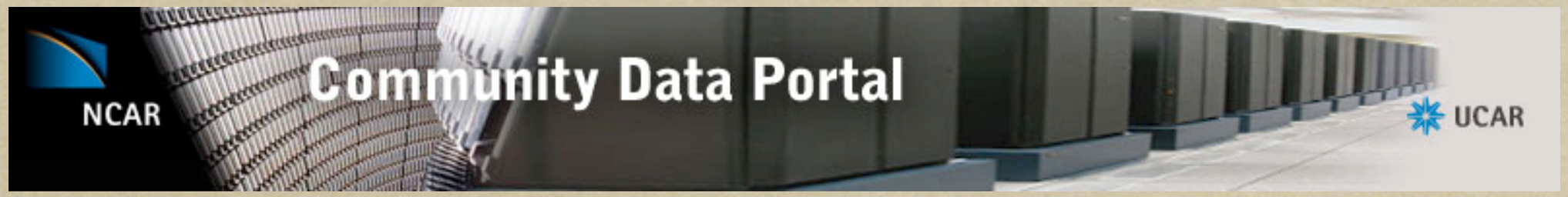
Emerging Cyberinfrastructure for Data at NCAR

Don Middleton
NCAR Computational and Information Systems Laboratory
Scientific Computing Division

June 6, 2005; GO-ESSP Gathering; UK Rutherford-Appleton Labs



CDP - Part of the NCAR Cyberinfrastructure Strategic Initiative, aimed at developing sustainable, effective strategies for data management and sharing

- *Establish integrated, shared* cyberinfrastructure *for managing and accessing geoscientific data: systems, middleware, applications*
- *Support many different projects and activities*
- *Start with institution, then broaden*
- *Build teams, partnerships, and technology*
- *Develop a portfolio of science and IT projects*

NCAR

Community Data Portal

UCAR

The Community Data Portal (CDP) is a collection of earth science datasets from NCAR, UCAR, UOP, and participating organizations in the following research areas:

- oceanic
- atmospheric
- space weather
- turbulence

CDP is an NCAR Cyberinfrastructure Strategic Initiative led by the NCAR Scientific Computing Division. It is a collaboration between UCAR, NCAR, UOP, and the National Science Foundation. CDP is developed by NCAR/SCD/VETS and utilizes several software packages and standards including LAS, OPeNDAP, GCMD, and THREDDS.

easy search | advanced search

Search for Earth Science datasets by metadata keyword:

Search

browse

- ACD/BAI/MEGAN : Model of Emissions of Gases and Aerosols from Nature
- ACD Model Evaluation Data : Merged data sets from aircraft campaigns
- ACD/MOZART Model : Model for OZone and Related chemical Tracers
- ATD/AIRS-II : Alliance Icing Research Study - II
- ATD/GOTEX : Gulf Of Tehuantepec EXperiment
- ATD/IHOP : IHOP 2002 campaign
- ATD/NAME : North American Monsoon Experiment
- CGD/CAS : Climate Analysis Section top level catalog
- CGD/CCSM : Community Climate System Model (data served by ESG)
- CGD/CDAS : Carbon Data-Model Assimilation
- CGD/DAYMET : Daily Surface Weather Data and Climatological Summaries
- CGD/PCM : Parallel Climate Model (data served by ESG)
- COLA : Center for Ocean-Land-Atmosphere studies
- CU/CIRES/ENLIL : Heliospheric Model

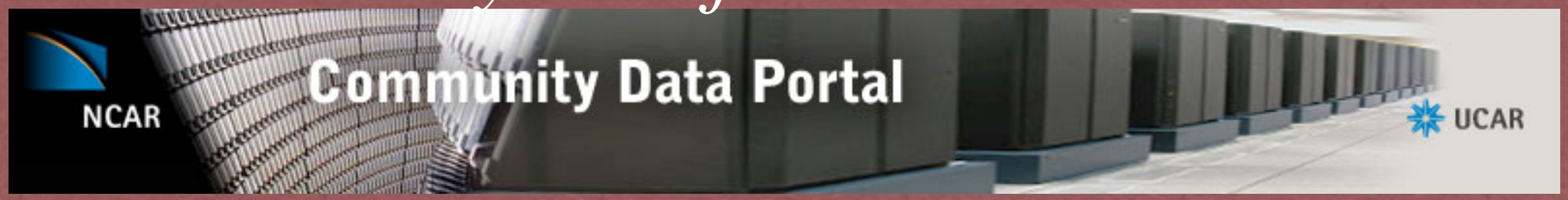
CDP: A Program

Projects & Relationships

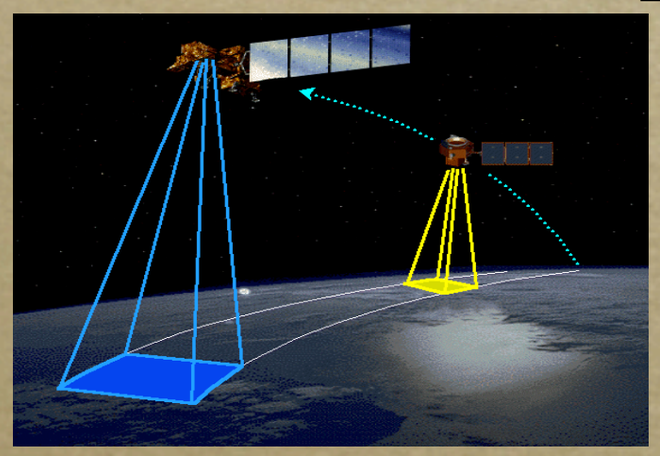
- *The Core CDP Portal*
- *The Earth System Grid (ESG)*
- *THREDDS (Unidata)*
- *GridBGC (NASA)*
- *Future Weather Information Systems (FWIS, WMO)*
- *Virtual Solar Terrestrial Observatory (VSTO)*
- *Chronopolis (in proposal stage with SDSC, Univ. of Maryland, NSF)*
- *British Atmospheric Data Center (BADC, Federation)*
- *GO-ESSP (Global Organization for Earth System Science Portals)*
- *And many scientific data partnerships: SCD/DSS, CCSM, PCM, WACCM, Unidata, ATD Aircraft/HIAPER, GIS Initiative, MEGAN, CISM, and many more*

Communities, Projects, and IT R&D

Cyberinfrastructure



Globus/Grid SRB FWIS LAS NCL
THREDDS GDS OPeNDAP OAI GIS



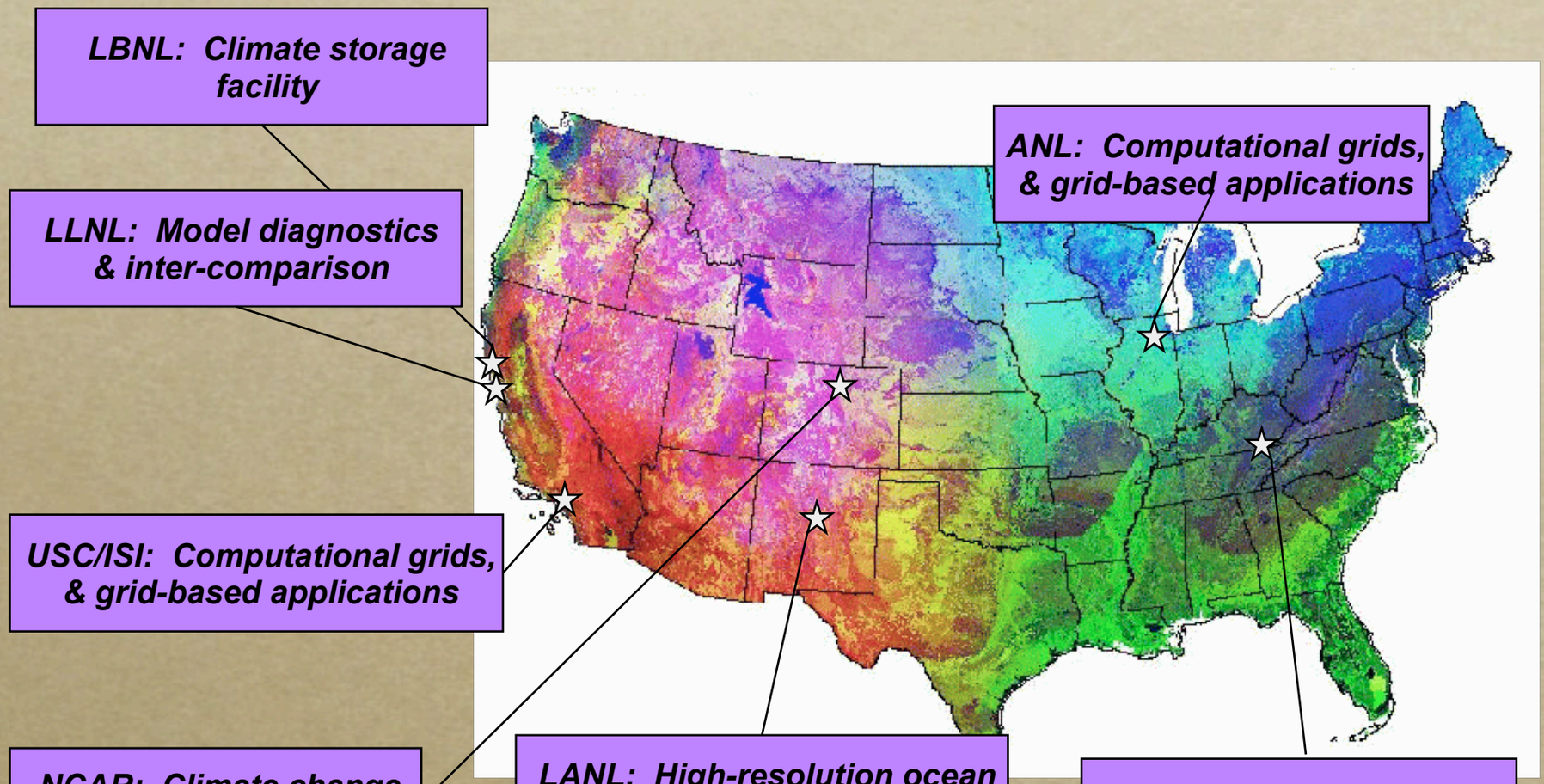
Models *Observation* *Archives* *Federation*

NCAR

NCAR Scientific Computing Division
Supercomputing • Communications • Data

NSF

The Earth System Grid (ESG)



http://www.earthsystemgrid.org

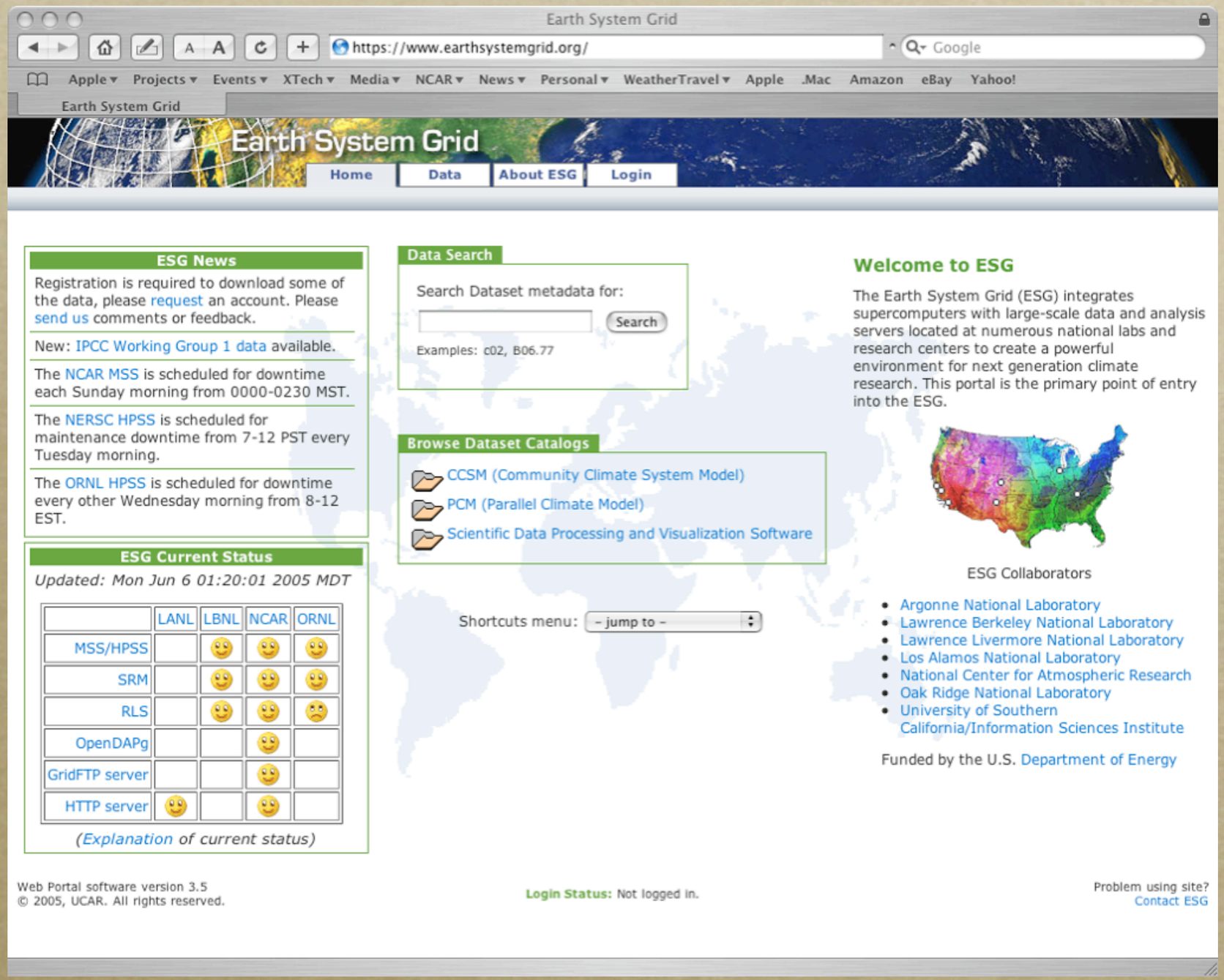
NCAR

NCAR Scientific Computing Division
Supercomputing • Communications • Data

NSF

Earth System Grid

An Operational DataGrid for Climate Research



NCAR

NCAR Scientific Computing Division

Supercomputing • Communications • Data

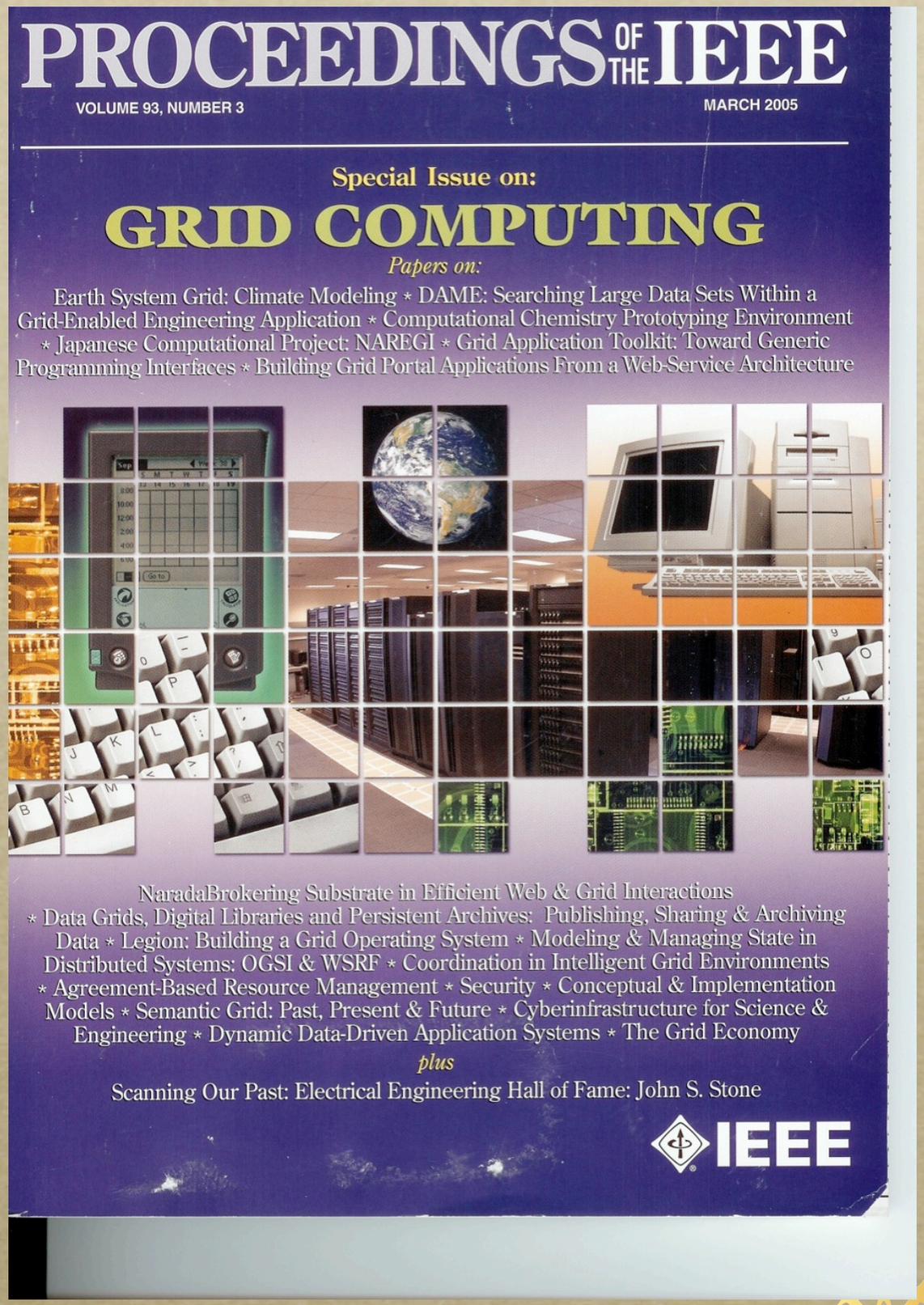
NSF

ESG Metrics

- *Two portals: US Climate Models and the Intergovernmental Panel on Climate Change (IPCC)*
- *CCSM/PCM Site: 620 registrations, 500 approvals, 578 datasets, 350K files, 50TB of data, 1.5TB's downloaded*
- *IPCC Site: 293 registrations, 44.3K files,18.7TB of data, 24.1TB downloaded*
- **Approx. Totals:** *800 registrations, 70TB of data, 400K files, 4 data sites, 26TB downloaded,* in 6-9 months of operation
- *Adding LANL and have proposal in with JPL PO-DAAC and ESMF*

ESG Paper

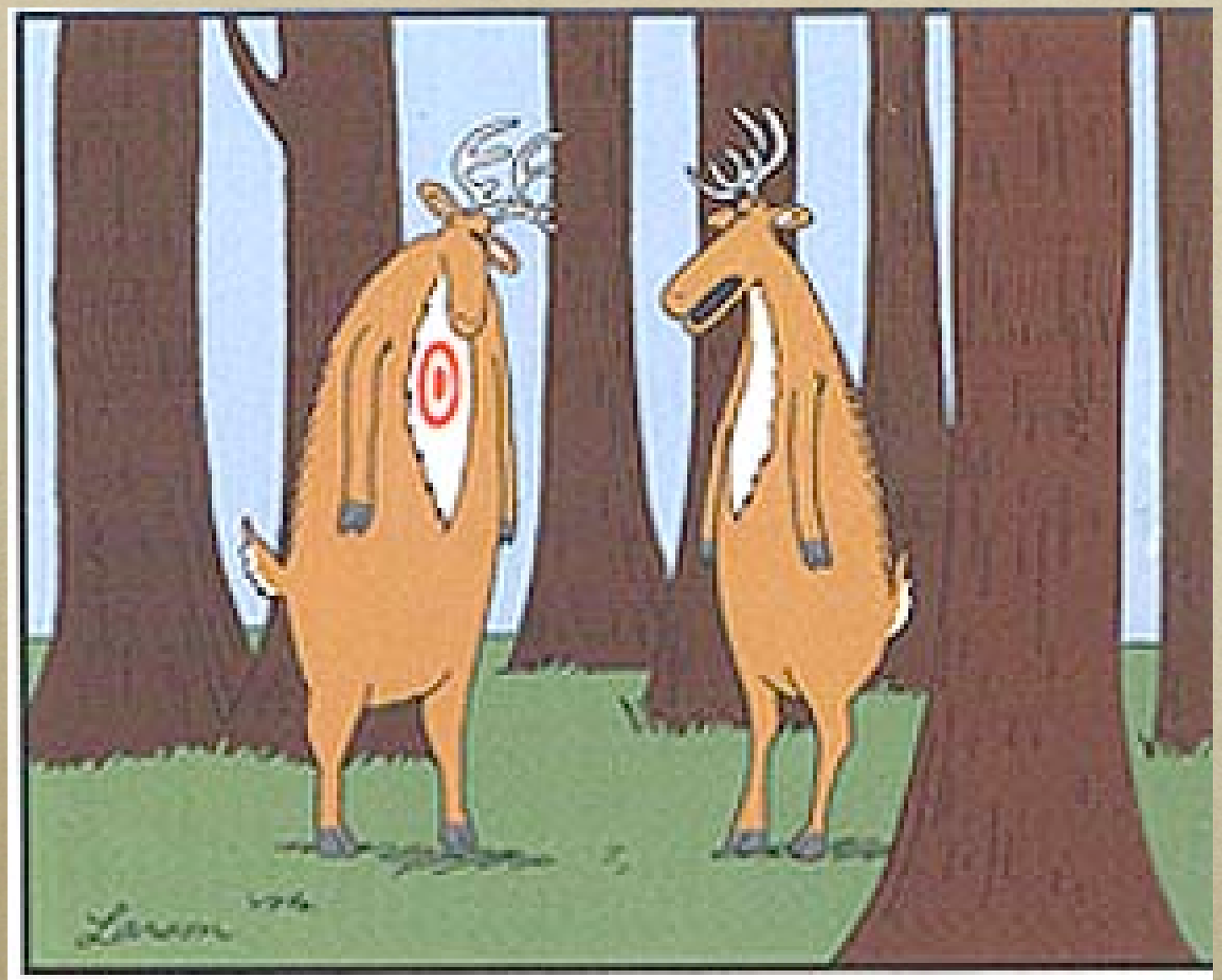
Proc. of the IEEE, March 2005



NCAR

NCAR Scientific Computing Division
Supercomputing • Communications • Data

NSF



"Bummer of a birthmark, Hal."

Security, the Final Frontier

The Virtual Solar Terrestrial Observatory (VSTO)

- VSTO is a collaborative project among
  - *NCAR's High Altitude Observatory*
  - *NCAR's Scientific Computing Division*
  - *Stanford's Knowledge Systems Lab*
- *VSTO is funded by a grant from the National Science Foundation, Computer and Information Science and Engineering (CISE) in the Shared Cyberinfrastructure (SCI) division.*

VSTO

The prototype Virtual Solar-Terrestrial Observatory (VSTO) is a distributed, scalable education and research environment for searching, integrating, and analyzing observational, experimental and model databases in fields of solar, solar-terrestrial and space physics.

- *Next generation data system aimed at integrating a dozen or so current generation ones*
- *Formal incorporation of knowledge/ semantics: ontologies, reasoning*

Chronopolis

- *A proposed collaborative effort of SDSC, University of Maryland, and NCAR*
- *Prototyping a petascale National Data Preservation Grid*
- *Appears to be poised for NSF Funding*

Earth System Curator

- *NSF Proposal, appears to be poised for funding*
- *A joint exploratory effort between ESMF and ESG*
- *Integrated approach to creating and using metadata that spans the model components, file-level syntactic metadata, and discovery level holdings*

Addressing the convergence of models and data

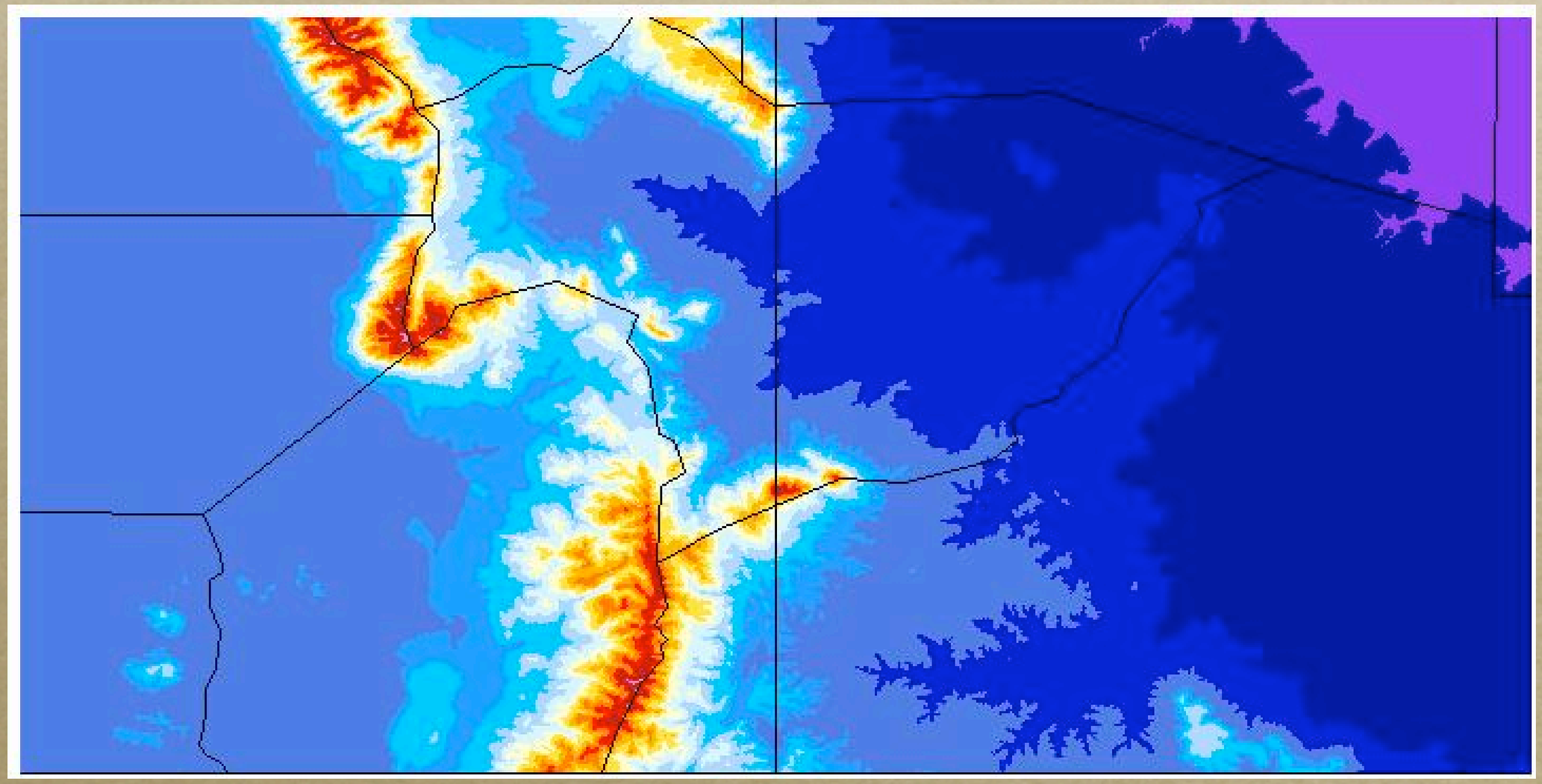
National Lambda Rail (NLR)

NCAR/SCD will serve: Colorado School of Mines, Colorado State University, University of Colorado at Boulder, University of Colorado at Denver, NOAA Boulder labs, University of Utah, and the University of Wyoming.

NLR provides 4 separate 10 gigabit-per-second pathways, extensible to ~40 total

Analysis & Visualization for the Geosciences

The NCAR Command Language (NCL)



NCAR

NCAR Scientific Computing Division

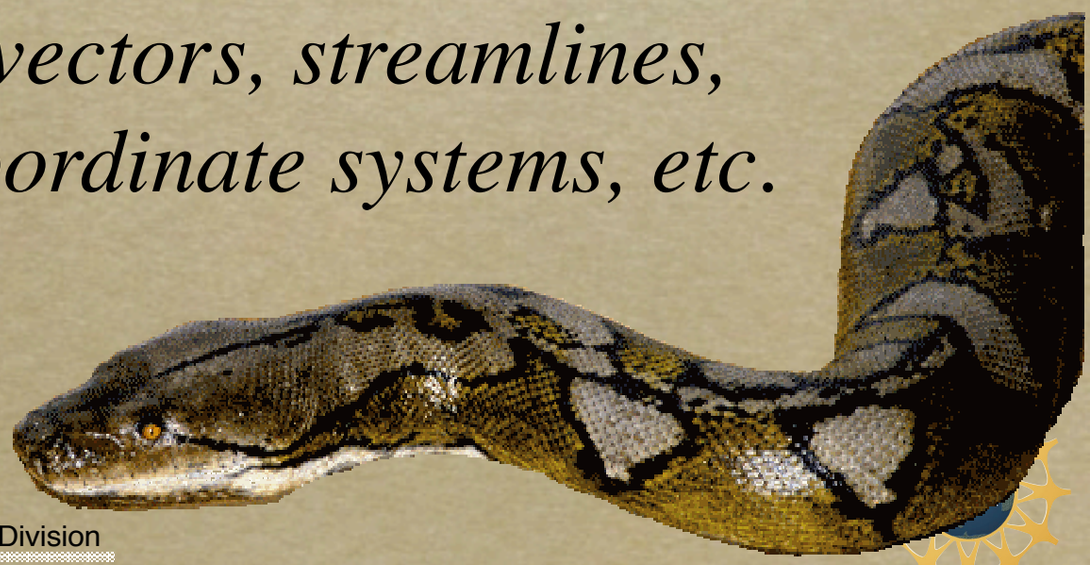
Supercomputing • Communications • Data

NSF

NCL

- *4th Generation Language (4GL's) tailored for geoscientific use*
- *Data Ingest: netCDF, HDF, GRIB, binary...*
- *Data Analysis: Over 400 functions*
- *Data Visualization: Contours, vectors, streamlines, comprehensive maps, native coordinate systems, etc.*

And now PyNGL



NCL/PyNGL

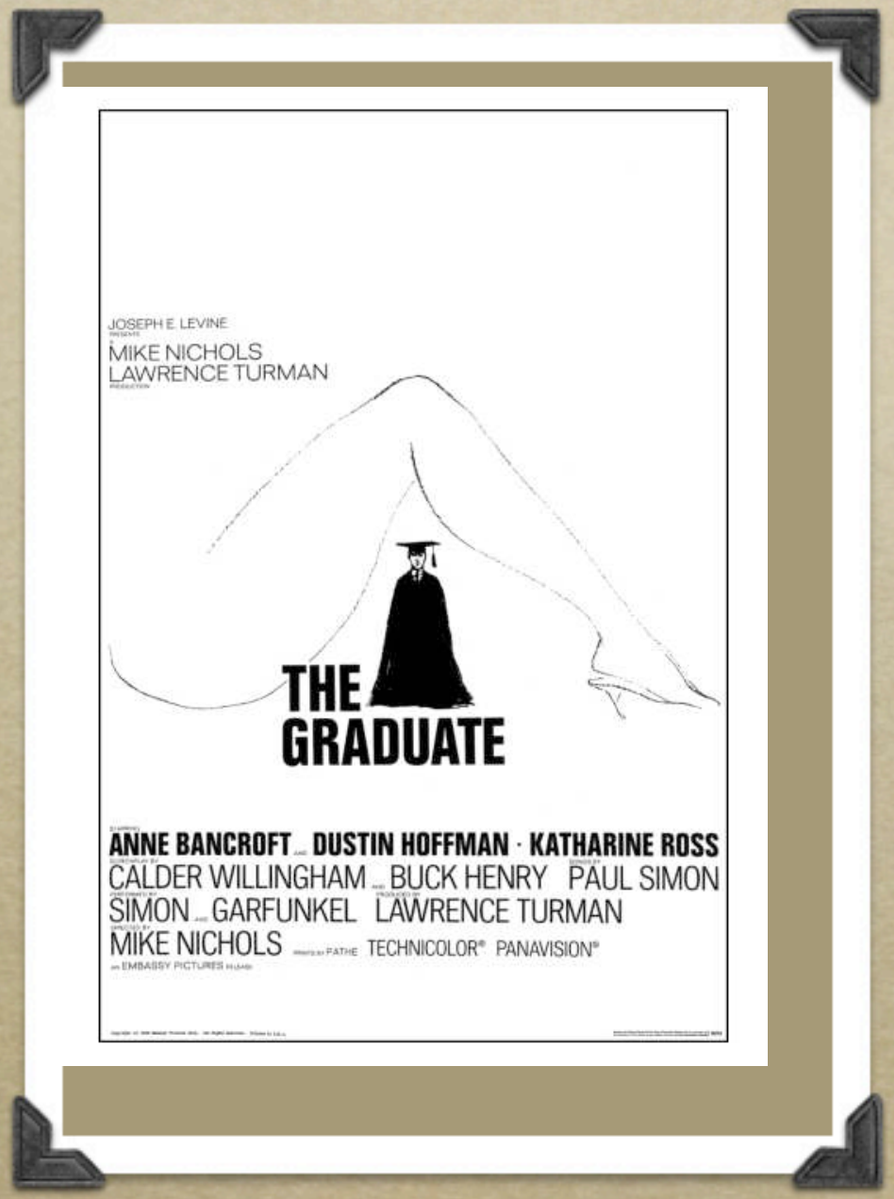
- *Development: 75+ FTE-years of effort*
- *Size:1.5 million lines of code, .5 million documentation*
- *User community: order 1000's*
- *Used: On supercomputers, by universities, DOE, DOD, NOAA, NASA, in the classroom, behind web*
- *On: Ports to all major operating systems (and then some): Unix, Linux, Mac OSX, Windows*
- *Successful Because: Serves community well, supported, documented, examples, training, free*

On Other Fronts…

- *Security concerns*
- *Deployment of SAN's, SNFS, and Commodity RAID for dataportals*
- *Beginning to weave Data Management into major science proposals*
- *U.S. funding to science agencies under pressure, metrics very important*

"I just want to say one word to you. Just one word..."

"Data"



1967